Engineering the acyltransferase substrate specificity of assembly line polyketide synthases

Briana J. Dunn\(^1\) and Chaitan Khosla\(^{1,2}\)

\(^1\)Department of Chemical Engineering, and \(^2\)Departments of Chemistry and Biochemistry, Stanford University, Stanford, CA, USA

Polyketide natural products act as a broad range of therapeutics, including antibiotics, immunosuppressants and anti-cancer agents. This therapeutic diversity stems from the structural diversity of these small molecules, many of which are produced in an assembly line manner by modular polyketide synthases. The acyltransferase (AT) domains of these megasynthases are responsible for selection and incorporation of simple monomeric building blocks, and are thus responsible for a large amount of the resulting polyketide structural diversity. The substrate specificity of these domains is often targeted for engineering in the generation of novel, therapeutically active natural products. This review outlines recent developments that can be used in the successful engineering of these domains, including AT sequence and structural data, mechanistic insights and the production of a diverse pool of extender units. It also provides an overview of previous AT domain engineering attempts, and concludes with proposed engineering approaches that take advantage of current knowledge. These approaches may lead to successful production of biologically active ‘unnatural’ natural products.

1. Introduction

Natural products have historically been an invaluable resource in the discovery and development of therapeutically active small molecules. Micro-organism-derived natural products became more prevalent following the discovery of penicillin and remained significant until the 1990s, when nearly 80 per cent of drugs were either natural products or natural product analogues [1]. Despite a subsequent decline in investments in natural product research by the pharmaceutical industry, the field is now better-equipped than ever to address critical therapeutic needs. Advancements in sequencing technologies and structural analysis techniques, combined with a better understanding of biosynthetic pathways and a wealth of unexplored source organisms, create an exciting environment for future drug discovery [1–3].

The polyketide class of natural products has seen huge success in the commercial drug arena, with a ‘hit rate’ that is several orders of magnitude above synthetic compound libraries [1]. Polyketides act as a wide range of therapeutics, including antibiotics (erythromycin A), immunosuppressants (FK506, rapamycin), anti-cancer agents (salinosporamide A, epothilone B) and others (figure 1). Many polyketides are produced by assembly line polyketide synthases (PKSs), which build polyketide chains by successive condensation of simple coenzyme A (CoA)-derived subunits. The structural and functional modularity of these enzyme complexes has placed them at the forefront of engineering endeavours in the pursuit of biologically active ‘unnatural’ natural products. A large amount of polyketide diversity stems from differences in building block selection by acyltransferase (AT) domains, and thus many engineering efforts have focused on altering the catalytic activity of these domains.

While the environment is right for major developments in polyketide engineering, these pursuits must be informed by previous efforts and new knowledge. This review begins by examining the characteristics of modular PKSs that originally placed them at the forefront of engineering efforts, and outlines recent developments that can be further used in the successful engineering of these enzymatic
assembly lines (figure 2). We examine previous attempts to engineer PKS substrate selection by AT domain engineering, such as AT domain swapping, AT knockout and complementation by a domain that normally functions in trans, and AT site-directed mutagenesis (figure 3). We conclude by proposing future engineering approaches that we believe will be most successful in the quest for novel polyketides. These structure and sequence-based approaches take advantage of the aforementioned information and build upon successful engineering of similar protein families. This will set the stage for productive drug development in the near future (figure 4).

Figure 1. The structural and therapeutic diversity of polyketide natural products. This diversity largely stems from differences in extender unit selection by the AT domain (for details, see text).

2. The acyltransferase domain as an engineering target

2.1. Polyketide diversity: the importance of the acyltransferase domain

Modular type I PKSs are found primarily in Gram-negative and -positive bacteria, and are especially prevalent in actinomycetes [4]. These megasynthases are comprised of several catalytic modules that are each responsible for one round of polyketide chain elongation. Each module is composed of
several different enzymatic domains, each of which plays a distinct role in the extension and modification of the growing polyketide. AT domains select the appropriate α-carboxyacyl-CoA building block and catalyse its loading onto the phosphopantetheine arm of the acyl carrier protein (ACP). Ketosynthase (KS) domains accept the polyketide chain from the upstream ACP and catalyse a Claisen-like condensation between it and the ACP-bound α-carboxyacyl-CoA extender.

**Figure 2.** Tools for engineering the substrate specificity of modular AT domains. A diverse pool of substrates, mechanistic insights and sequence and structural information combine to allow for productive AT substrate specificity engineering using a wide variety of techniques (for details, see text). (Online version in colour.)
Figure 3. Previous attempts to engineer AT domain substrate specificity. The ‘acceptor’ AT domain is defined as the starting engineering scaffold, whereas the ‘donor’ AT domain possesses the final desired substrate specificity and is used to either fully replace or provide a template for mutation of the acceptor domain. MMCoA, methylmalonyl-CoA; MCoA, malonyl-CoA; EMCoA, ethylmalonyl-CoA; MeO-ACP, methoxymalonyl-ACP; PPM-SNAC, 2-propargylmalonyl-N-acetylcysteamine. (Online version in colour.)
unit. Together, the KS, AT and ACP comprise the minimal module that is necessary and sufficient for one round of chain elongation. Accessory domains such as dehydratase (DH), enoyl reductase (ER) and keto-reductase (KR) domains may also be present in a module and control the level of oxidative-reductive modification of the β-keto group following chain elongation. The thioesterase (TE) domain is responsible for release of the final polyketide product, either by cyclization or by attack of an external nucleophile.

Modular PKS genes and their corresponding proteins are co-linear [5,6]—the organization of the assembly line (and thus the structure of the polyketide product) can often be predicted from the gene sequence. This combined modularity has made these catalysts attractive targets in the engineering of novel products via combinatorial biosynthesis [7,8]. Most domains in the assembly line receive their substrates from previous catalytic steps and not by diffusive loading; these enzymes therefore may not have been evolutionarily driven to develop strict substrate specificities and could be inherently robust to rearrangement [9]. Indeed, loading AT domains, KS, ACP, KR, DH and ER domains from the 6-deoxoerythronolide B synthase (DEBS) have been shown to accept substrates that vary considerably from their native substrates [10–15]. Alternatively, extender AT domains receive their substrates diffusively and usually exhibit strict specificity towards a single α-carboxyacyl-CoA building block [10,14]. In acting as the primary gatekeepers to building block incorporation, these domains play an important role in polyketide diversification and are often the primary target of engineering efforts. They are also the focus of this review.

### 2.2. Extender unit diversity

Extender AT domains in modular PKSs are naturally specific for several different α-carboxyacyl-CoA substrates, but malonyl-CoA and methylmalonyl-CoA are most commonly used in the biosynthetic pathways characterized to date [16]. AT domains have also been identified that are capable of recognizing ACP-bound substrates, such as those involved in the biosynthesis of geldanamycin [17], zwitsermicin A [16,18–20] and guadinomine [21]. In vivo and in vitro studies aimed at interrogating AT substrate specificity have previously been limited by the cost and availability of these extender units [22–23], but methods for producing diverse pools of extender units enzymatically have emerged in recent years.

The discovery of crotonyl-CoA carboxylase/reductase (CCR) homologues in polyketide biosynthetic pathways more than doubled the number of CoA-linked extender units known to be incorporated into polyketides. These enzymes catalyse the reductive carboxylation of α,β-unsaturated acyl-CoA precursors (for review, see Wilson & Moore [24]). Originally discovered in the context of ethylmalonyl-CoA [25], CCR functionality quickly expanded to include biosynthesis of alkylmalonyl-CoA and both straight and branched chain malonyl-CoA derivatives. The AT domains associated with CCR-generated extender units typically show relaxed specificity, and several polyketide analogues have been made that take advantage of this fact, but more ambitious endeavours will benefit from rational engineering [24].

Malonyl-CoA can be produced by the direct activation of malonate by an ATP-dependent malonyl-CoA synthetase (MatB; [26,27]). The malonyl-CoA synthetases from Rhizobium trifolii [28] and Streptomyces coelicolor [22] were shown to have broad substrate tolerance, activating several different 1,3-dicarboxylylate substrates with measurable efficiency. The S. coelicolor MatB was also shown to be promiscuous towards thiol acceptors, and is capable of producing N-acetylglutamine- (SNAC-) and pantetheine-linked extender units [22]. AT domains have been shown to accept SNAC and pantetheine analogues, and these substrates are readily taken up by cells from exogenous media [22,29–35]. The activity of the R. trifolii MatB towards several natural and unnatural malonate derivatives was enhanced using structure-guided saturation mutagenesis [36,37]. Interestingly, the trans-AT domain from the kirromycin PKS was shown to be more promiscuous than originally anticipated when examined for transacylation activity with MatB-generated extender units [37].

A diverse pool of CoA-linked extender units provides a valuable toolbox for the polyketide engineer. The ability to regiospecifically incorporate many of the possible MatB or CCR-derived CoA-linked extender units into DEBS, for example, has the potential to yield over one million novel products alone. Availability of these extender units for in vivo and in vitro studies allows for interrogation of the structural and mechanistic determinants of substrate specificity. The emergence of AT domains capable of selecting substrates not present in the native host also provides new starting scaffolds in the pursuit of AT domains with altered specificity.

### 2.3. Acyltransferase mechanistic insights

AT-catalysed incorporation of CoA-derived extender units occurs by a ping–pong bi–bi mechanism [38,39] involving an acyl-AT intermediate that is subject to nucleophilic attack by the thiol residue found on the phosphopantetheine arm of the ACP. Early studies revealed that modular AT domains likely do not possess epimerase activity, and all domains in DEBS exclusively incorporate (2S)-methylmalonyl-CoA [40]. The rapid rate at which AT domains catalyse transacylation has made kinetic characterization of their activity apart from the rest of the assembly line difficult. Recently, the AT domain from module 3 of DEBS was characterized using a coupled enzymatic assay that allowed for continuous monitoring of AT catalytic activity and thus characterization of α-carboxyacyl-CoA and ACP specificity. This study revealed that the (2S)-methylmalonyl-CoA specificity of this enzyme lies in the first half reaction of the ping–pong mechanism (formation of the methylmalonyl-AT intermediate; [39]), Hydrolytic cleavage of incorrect substrates has been hypothesized to be a specificity-determining mechanism used by AT domains, but the reported importance of this mechanism has varied [39,41].

Several studies have established the importance of proper AT–ACP interactions during transacylation. Examination of domains from modules 3 and 6 of DEBS revealed that cognate AT–ACP pairs have at least 10-fold greater specificity for each other than for heterologous proteins [39,42]. Polyketide chain translocation [43] and elongation [15] are also dependent on proper protein–protein interactions within a module, and specific portions of the AT domain have been identified that contribute to these interactions. For example, the cleft formed by the KS domain, KS to AT linker, and AT domain was shown to be a possible docking site for the ACP during chain transfer and elongation [44,45]. Conservation of
2.4. Acyltransferase domain sequence and structure

The development of high-throughput sequencing technologies and a rise in whole-genome sequencing has led to a rapid increase in the number of PKS cluster sequences available [46]. Hundreds of AT domains have been sequenced and their clusters biochemically characterized [47], providing a useful dataset in the engineering of these domains using statistical, sequence-based methods. Many AT domains share high sequence homology and can be identified computationally based strictly on their sequence similarity to a template AT domain [48]. Interestingly, divergence occurs on the basis of extender unit specificity and not species of origin or composition of the rest of the PKS module [49]. Phylogenetic analyses of AT domains annotated for substrate specificity produce two major clades based on methylmalonyl- or malonyl-CoA specificity [48–50], and this separation is improved when considering only active site and highly conserved residues that vary with methylmalonyl- or malonyl-CoA substrate specificity [48]. Residues important in the prediction of ethylmalonyl- and methoxymalonyl-CoA specificity have also been identified [4], and the prediction of less common specificities will likely become more accurate as more sequenced AT domains are functionally characterized.

The X-ray crystal structures of two modular AT domains from DEBS have been solved (Protein Data Bank (PDB) entries 2QO3 and 2HG4) [51,52], as well as a trans-acting AT domain from the disorazole synthase (PDB entry 3RGI) [53]. Along with the multi-purpose FabD AT domains from *Escherichia coli* (PDB entry 1MLA) [54] and *S. coelicolor* (PDB entry 1NM2) [55], these crystal structures have facilitated the identification of residues that may play a role in determining substrate specificity. The *E. coli* FabD protein was crystallized with malonyl-CoA (PDB entry 2G2Z), allowing for the identification of conserved oxanion hole residues (Gln11, Leu93; 2G2Z numbering), and a conserved arginine residue (Arg117; 2G2Z numbering) that interacts with the carboxylate of the CoA-linked extender unit [56]. Crystal structures have also been useful in conjunction with multiple sequence alignments to identify residues in and around the active site that show strong correlations with substrate specificity, including four residues that have been identified repeatedly in different contexts (Leu93, Ser197, Val198, Ser200; 2G2Z numbering) [48,57–60]. The structure-based sequence analysis of polyketide synthases server allows for three-dimensional modelling of AT domains that have not been crystallized, and also allows for sequence-based comparisons and prediction of inter-subunit contacts [47].

Additionally, an AT domain with atypical specificity for long-chain acyl-CoA substrates from the mycobacterial PKS Pks13 (PDB entry 3TZX) was also crystallized recently, revealing a dedicated channel for the atypical long chain and providing an interesting comparison for previously crystallized methylmalonyl- and malonyl-CoA AT domains [61]. The structural advances made in elucidating AT domain architecture, as well as progress in describing the three-dimensional conformation of the rest of the modular PKS [62] position the protein engineer to pursue many different structure-based engineering approaches.

3. Previous attempts to engineer acyltransferase substrate specificity

3.1. Acyltransferase domain swaps

Despite the identification of specific residues that probably play a role in determining substrate specificity, the most common approach for altering AT catalytic activity to date is the swapping of an entire AT domain for a homologue with different specificity. Several desmethyl erythromycin, desmethyldeoxyerythronolide B (desmethyl-6-deoxyerythronolide B) or desmethyl triketide lactone analogues have been successfully produced by swapping methylmalonyl-CoA specific domains in DEBS for malonyl-CoA specific AT domains from the rapamycin PKS [40,63–67]. A 6-desmethyl-6-ethylnolactone A analogue was produced by swapping AT4 of DEBS with an ethylmalonyl-CoA specific AT domain from the niddamycin PKS. When expressed in *Saccharopolyspora erythraea*, the hybrid PKS produced only erythromycin A unless the culture media was supplemented with precursors to ethylmalonate. The 6-desmethyl-6-ethylnolactone A analogue was the primary macrolide produced when the *cr* gene from *Streptomyces collinus* was also expressed in this strain [68]. A methoxymalonate building block was also incorporated into the 6-DEB scaffold to form 2-desmethyl-2-methoxy-6-DEB by swapping AT6 of DEBS with AT8 of the FKS20 PKS [69]. AT domains in the geldanamycin PKS specific for methylmalonyl- or methoxymalonyl-CoA were successfully replaced with malonyl-CoA specific domains from the rapamycin PKS. These domain swaps led to the identification of a geldanamycin analogue with fourfold higher affinity towards Hsp90, a protein drug target [70]. Domain swaps have been successfully implemented with other PKSs as well [59,64,71].

Despite the apparent success of AT domain swaps in producing novel polyketide products, these successes often come at the cost of decreased product titres [72], and some AT domains appear to be especially difficult to replace [58]. While the causes of such shortcomings are not well understood, disruption of proper protein–protein interactions or an inability of downstream modules to handle altered substrates may be to blame [73]. Titres are decreased to varying degrees depending on the location of the swap in the PKS assembly line [40,64], indicating variability in the ability of downstream modules to process non-native substrates. The extensive interdomain interface shared by AT domains with their KS neighbours [51,52] could also be disrupted by domain swapping, as could the interactions of the ACP with the KS–AT didomain during chain elongation [44,74].

3.2. Generation of hybrid acyltransferase domains

A less common approach to altering AT catalytic activity involves creation of hybrid AT domains, in which possible specificity-determining regions are replaced by the analogous cassettes from an AT domain with a different α-carboxyacyl-CoA specificity. Hybrid AT domains created using methylmalonyl-CoA specific domains from DEBS and the malonyl-CoA specific AT domain from module two of the rapamycin PKS suggested that a short C-terminal ‘hypervariable region’ was the primary determinant of AT specificity [40]. The importance of this region in the context of other AT domains is not clear, but it is important to note given the success of hybrid domains in altering the reactivity of other modular PKS domains [44,74].
3.3. Acyltransferase site-directed mutagenesis

As mentioned above, examination of multiple sequence alignments and crystal structures has led to the identification of several residues in and adjacent to the AT active site that diverge based on α-carboxyacyl-CoA substrate specificity [48,57–60]. Site-directed mutagenesis of important residues has the potential to be minimally invasive, allowing the AT domain to remain in its native environment and causing minimal perturbation to important protein–protein interactions. A motif approximately 100 residues C-terminal of the active-site serine is commonly targeted for mutagenesis; mutation from the methylmalonyl-CoA specific YASH motif to the malonyl-CoA specific HAFH motif in AT1 [59], AT4 [58] and AT6 [60] of DEBS consistently led to promiscuous AT domains capable of incorporating both extender units. Mutations outside of this motif have also led to promiscuity, but have allowed incorporation of the non-native extender unit to a lesser extent [58,60]. Molecular dynamics (MD) simulations followed by Quantum Mechanics/Molecular Mechanics (QM/MM) optimizations on DEBS AT6 allowed for the identification of a residue in the active-site pocket that appeared to constrain the size of the extender unit side chain. Mutation of this residue from valine to alanine allowed for incorporation of the non-natural substrate 2-propargylmalonyl-SNAC, although at limited quantities [60].

It is important to note that alterations in AT substrate specificity are often quantified in vivo, and thus the availability of different α-carboxyacyl-CoA substrates cannot be readily controlled. The relative kinetic parameters describing the specificity of a mutated AT domain cannot be determined using in vivo techniques, and until recently the kinetic effects of such mutations were not known. In vitro analysis of a YASH to HAFH DEBS AT3 mutant, along with another point mutant previously examined in vivo [58,60] revealed that the activity of these mutants is drastically attenuated, implying that incorporation of non-natural extender units likely occurs as a result of diminished AT catalytic activity and not because of increased specificity towards the non-native substrate [39]. The engineering of AT domains by site-directed mutagenesis therefore has vast room for improvement and will benefit greatly from the use of advanced sequence- and structure-based techniques.

3.4. Trans-acyltransferase complementation

Several modular PKSs have been identified that lack canonical cis-AT domains and instead use freestanding domains that act in trans, similar to freestanding type II fatty acid synthase AT domains [75]. These trans-AT PKSs include those responsible for production of the disorazoles [76,77], kirromycin [78–80] and several others [75,81,82]. A potential engineering strategy involves the selective inactivation of a cis-AT domain and complementation using a trans-AT with differing specificity [9]. This approach has the potential to impose minimal conformational strain on the synthase. The majority of trans-AT domains are thought to be specific for malonyl-CoA, but domains with varying specificities have been proposed recently [80]. The ethylmalonyl-CoA specific kirromycin trans-AT domain was shown to be promiscuous, capable of catalysing the transacylation of several α-carboxyacyl-CoA substrates not found in its native host [37]. Further, several trans-AT domains have shown promiscuity towards their ACP substrates as well, with activities towards heterologous ACP domains that meet or surpass the activities of native cis-AT domains [42,83]. Several examples exist of trans-AT complementation in the context of an AT-null PKS module. The malonyl-CoA:ACP transacylase from S. coelicolor was used with an AT-null DEBS module 6 in the production of 2-desmethyl-t-6-dEB [83] and the trans-AT domains from the bryostatin PKS also complemented an AT-null DEBS module 6 [84].

In engineering cis- and trans-AT domains, it is important to note that these domains differ in their evolutionary origins. While cis domains are found mostly in actinomycetes, myxobacteria and cyanobacteria, trans domains stem largely from proteobacteria, myxobacteria and bacilli [75]. These domains appear to have evolved independently; cis-AT domains appear to have evolved largely by gene duplication and diversification, whereas trans domains evolved largely by horizontal gene transfer [49,75,81,85]. Trans-AT PKSs exhibit diverse modular architectures and deviate from normal colinearity rules [81], and the KS domains of these PKSs have been used to predict product structure. Phylogenetic separation of these domains is largely based on substrate structure, and PKS module function and overall product structure can be deduced from the predicted substrate specificity of each KS domain [75,86,87].

4. The future of acyltransferase domain engineering

AT domain engineering by a variety of computational and experimental techniques is now possible due to the availability of high-quality sequence and structural data and advancements in in vitro assay techniques. Specifically, approaches in both directed evolution and rational design—hallmarks of protein engineering in the past several decades—can be pursued given the advancements in AT domain architecture and function outlined above. Future engineering of AT domain specificity will require consideration of protein–protein interactions and the effects of alterations on each step of the catalytic mechanism. We outline below several techniques that may be especially relevant for engineering AT domain specificity.

4.1. Amino acid coevolution

Given the extensive sequence data available for modular AT domains, techniques using multiple sequence alignments are particularly useful. Previous utilization of AT alignments has been limited to identification of partially conserved residues based on substrate specificity [48,57–60]. These residues were identified using phylogenetic trees in combination with multiple sequence alignments in order to identify patterns of sequence separation. Owing to the limited success of targeted mutagenesis attempts based on these analyses, it is apparent that a fundamental driving force for specificity is not being captured using conservation alone. The natural next step is incorporation of coevolutionary information, or the correlated mutation of residues in response to an evolutionary pressure. We speculate that, beyond the highly conserved motifs already identified in AT domains, there are other residues that have evolved in a compensatory manner. Several methods have been developed for analysing amino acid coevolution in protein sequences (see reviews [88–90]). We focus here on a few selected techniques that
have been used successfully to alter or interrogate the substrate specificity of other protein families.

Residues that are differentially conserved based on subfamily, such as the residues that vary based on methylmalonyl- versus malonyl-CoA specificity, are often referred to as ‘specificity-determining positions’. It is assumed that these residues must coevolve because they show correlated mutational patterns [88]. Several computational techniques can be used to identify groups of coevolving residues in large multiple sequence alignments [91–101]. The JDet software package was developed recently for identification and manipulation of coevolving residues from an input alignment using several different methods [102]. One method, referred to as S3det, makes use of multiple correspondence analysis, which allows for extraction of independent sources of variation in the alignment. These variations can be linked to corresponding clusters of proteins in order to identify subfamilies. S3det was used to successfully identify ligand-binding sites in several protein families from the Pfam database [100]. As another example, the substrate specificity of a plant acyl-ACP TE was engineered using specificity-determining positions identified via a conserved property difference locator, which examines differential conservation as well as such consensus properties as size and hydrophobicity [96,103]. The use of computational techniques to identify specificity-determining positions could improve AT domain engineering, which may have previously been limited by sequence sample size and often relied on manual inspection or examination of crystal structures to identify specificity-determining residues.

Statistical coupling analysis (SCA) has been used on several occasions to identify groups of residues that have coevolved for the determination of substrate specificity. SCA goes a step beyond previously described methods by explicitly accounting for amino acid covariance. It uses a ‘SCA matrix’, or a conservation-weighted position correlation matrix that describes mathematically the conservation at individual amino acid positions and their inter-residue covariance [104,105]. Recent implementations of SCA identify statistically significant correlations among groups of residues, or ‘sectors’, by spectral decomposition of the SCA matrix followed by principal component analysis. This approach also allows for examination of the relationships between sequence divergence and amino acid covariance using a conservation-weighted sequence correlation matrix [106,107]. One of the first implementations of SCA successfully identified residues important for the substrate specificity of SIA serine proteases, and mutation of sector residues clearly perturbed the relative specificity constants ($k_{\text{cat}}/K_{m}$) for different substrates [105]. SCA was also used in the design of unnatural WW domains, small peptide recognition motifs that are classified based on the nature of their proline-containing target motif [108,109]. A group of eight residues that showed a high degree of coevolution was proposed to encode the binding specificity of these domains. These eight coevolving sector residues showed patterns of conservation that were strongly correlated with motif specificity [109]. SCA can identify residues that are not immediately obvious using conservation or structure examination alone and thus has the ability to expand and improve upon the cohort of previously identified AT specificity-determining residues.

The functional importance of residues identified using coevolution techniques depends on the constraints driving the evolution of the protein family of interest. The clear phylogenetic separation of AT domains into clades based on $\alpha$-carboxyacyl-CoA substrate specificity provides evidence that evolutionary pressure likely exists for this function. Coevolution analyses on modular AT domains will therefore likely lead to the identification of previously elusive residues that, when modified, will lead to alterations in specificity. Further, coevolution analyses have been used to detect protein interaction surfaces (see reviews [88–90]), and thus may be helpful in identifying residues necessary for the protein–protein interactions that allow for proper functioning of the assembly line. Knowledge of the interaction surfaces between AT domains and other proteins in the synthase could enable engineering of specificity that properly maintains these interactions, and could also shed light on the proper junctions for engineering based on AT domain replacement.

### 4.2. Structure-based techniques

The availability of several X-ray crystal structures of AT domains provides the opportunity for utilization of structure-based engineering. These crystal structures have provided guidance in previous AT engineering attempts [48,57–60], but have mostly been used to pinpoint residues that might make contact with the substrate. More global approaches could allow for more successful engineering of AT specificity.

Understanding of the molecular basis for AT substrate specificity would be enhanced by MD simulations, which have largely been missing from modular PKS studies until recently [60,62,110,111]. MD simulations rely on Newtonian physics to approximate molecular interaction and conformation energetics and to model atomic movement [112–114]. A structural model of DEBS AT6 with (25)-methylmalonyl-CoA docked into the active site was examined using a 30 ns MD simulation followed by QM/MM optimization. This simulation confirmed the role of proposed active-site residues and pinpointed residues important in maintaining the proper chemical environment for polar and non-polar regions of the substrate. MD simulations on mutated systems with the native methylmalonyl-CoA and non-native malonyl-CoA helped provide rationale for the observed in vivo substrate specificity of these mutant AT6 domains. Further, the MD simulations helped pinpoint a residue that appeared to prevent the incorporation of substrates with larger $\alpha$-substituents. Mutation of this residue allowed for detectable utilization of a 2-propargylmalonyl-SNAC building block [60]. The utility of this initial MD simulation on a modular AT domain provides motivation for the future implementation of this technique. Allowing MD simulations to play a more active role in the design of experiments, as opposed to using them primarily to explain observed phenomena, will likely aid in the identification of residues outside the active site that indirectly affect active-site architecture. Examination of long-time dynamics in the presence of native and non-native substrates will be a likely strategy in the pursuit of this goal.

Other QM-based techniques have also been used to examine enzymatic substrate specificity. Adenylation domains from non-ribosomal peptide synthetases, which are analogous to AT domains in PKSs, have been engineered to take up different amino acid substrates using a structure-based algorithm known as $K^*$. $K^*$ is an ensemble-based computational algorithm that models protein–ligand binding using approximate partition functions and energy minimization [115].
phenylalanine-specific adenylation domain from the gramicidin synthase was first engineered for improved specificity towards leucine residues [116], and after several improvements to the K* algorithm [117–119], was engineered for a complete specificity swap to leucine and improved specificity towards several other amino acids [120].

Finally, the engineering of AT domains using structural techniques would be further improved by the structural elucidation of AT domains with other α-carboxyacyl-CoA specificity. With the exception of the AT domain from the mycobacterial Pks13 [61], all modular AT domains crystallized to date are specific for methylmalonyl- or malonyl-CoA [51–55]. Examination of the active-site architecture and global structural characteristics of domains with more unique specificities, along with simple structural alignments with currently elucidated structures, could shed light on the structural determinants of building block specificity. Further, the ACP specificity of AT domains and the role this specificity plays in allowing for productive transacylation could be elucidated via the crystallization of an AT–ACP complex. Structural characterization of these complexes has proven difficult, likely due to the naturally dynamic nature of the ACP domain. Several methods exist for the cross-linking of ACP and KS domains based on modification of the prosthetic phosphopantetheine group on the ACP [121,122], but no such methods exist for AT–ACP cross-linking. A method was devised for cross-linking and subsequent purification of the trans-AT from the disorazole PKS with its cognate ACP from module 1 with the intention of using this adduct as a crystallography tool. This method used an active-site

<table>
<thead>
<tr>
<th>approach</th>
<th>example analytical tool</th>
<th>example application</th>
</tr>
</thead>
<tbody>
<tr>
<td>amino acid coevolution</td>
<td>S3det [94]</td>
<td>ligand-binding sites of several protein families in Pfam database [94]</td>
</tr>
<tr>
<td></td>
<td>conserved property difference locator [90]</td>
<td>substrate specificity of a plant acyl-ACP thioesterase [97]</td>
</tr>
<tr>
<td>statistical coupling analysis</td>
<td>statistical coupling analysis [98–101]</td>
<td>substrate specificity of S1A serine proteases [99]</td>
</tr>
<tr>
<td>structure-based techniques</td>
<td>molecular dynamics simulations [106–108] and quantum mechanics/ molecular mechanics optimization</td>
<td>substrate specificity of DEBS AT6 [54]</td>
</tr>
<tr>
<td></td>
<td>K* algorithm [109,111–113]</td>
<td>substrate specificity of adenylation domains from non-ribosomal peptide synthetases [110,114]</td>
</tr>
<tr>
<td></td>
<td>protein cross-linking for subsequent crystallization [47]</td>
<td>AT–ACP interactions [47]</td>
</tr>
<tr>
<td>directed evolution</td>
<td>error-prone PCR and saturation mutagenesis followed by whole- cell screening</td>
<td>substrate specificity of LovD, a tailoring AT domain involved in lovastatin synthesis [121]</td>
</tr>
</tbody>
</table>

Figure 4. The future of AT domain engineering. Given the existing tools for engineering AT domain substrate specificity, several different techniques are feasible. The identified techniques have previously been used to engineer the substrate specificity of other protein families. (Online version in colour.)
serine to cysteine AT mutant, which was cross-linked to the phosphopantetheine arm of the ACP using dibromopropionone. Purification of the complex from free AT and ACP proteins was accomplished using a two-step affinity chromatography procedure [53]. Crystallization of this complex and others like it would be useful in understanding more fully the protein–protein interactions between AT and ACP domains, thus allowing for preservation of these domain interactions in engineered systems.

4.3. Directed evolution

Advancements in analytical techniques capable of rapidly analysing AT catalysis [39] make directed evolution of the AT domain an experimental possibility, although testing of a very large number of variants is still not technically feasible. Directed evolution techniques (for reviews, see [123–125]) rely on the creation of a mutant library by such methods as error-prone polymerase chain reaction (PCR), DNA shuffling or saturation mutagenesis. This library is then screened for improved function, and the identified mutants are subject to another round of mutagenesis and selection in a process that mimics natural evolution. This process is usually repeated several times until the desired function is achieved. Screening of all possible mutants is not feasible or desirable, and several in silico techniques have been developed to help design the most optimal mutant library [126]. While not necessary, computational optimization is helpful when an extremely high-throughput assay does not exist.

The substrate specificity of a tailoring enzyme involved in converting monacolin J acid (MJA) into the cholesterol-lowering drug lovastatin was engineered using directed evolution techniques. LovD, an AT that naturally transfers α-S-methylbutyrate to MJA was engineered for improved activity towards α-dimethylbutyrate, a substrate that when added to MJA creates the blockbuster cholesterol-lowering drug simvastatin. Mutant libraries were created by either error-prone PCR or saturation mutagenesis and screened using a whole-cell growth inhibition assay. Seven rounds of mutagenesis and screening produced a variant with six point mutations that exhibited approximately 11-fold higher whole-cell activity towards simvastatin synthesis and increased thermal stability. Crystallization revealed that mutated residues were scattered on the tertiary structure, distant from the active site and in both solvent-exposed and buried regions. These residues, like most residues identified using directed evolution, would have been difficult to identify using the crystal structure alone [127].

5. Conclusions and future outlook

The potential for engineering PKSs to produce novel bioactive molecules is essentially limitless. These molecular assembly lines lend themselves well to engineering, as the co-linearity of genes and gene products along with the predictability of the corresponding polyketide allow for rational manipulation with an intended product in mind. AT domains act as gatekeepers to polyketide extender unit incorporation and thus control a large source of the structural diversity of these natural products. The field is better-equipped than ever for AT domain engineering, with mechanistic, sequence, and structural knowledge on this domain growing at a fast pace in recent years. Many engineering techniques are now open to the AT domain engineer. It is likely that no single method will be entirely effective in altering the specificity of AT domains, and success will come with combinations of these techniques. Bridging of in vitro results to in vivo production environments will be important for the successful production of novel therapeutics. Further functional and structural characterization of sequenced AT domains, along with better tools for manipulating, expressing and analysing both individual domains and the overall PKS will continue to aid in making the production of novel polyketides a reality. The engineering of AT domain catalysis requires an in-depth knowledge of how such alterations affect the entire polyketide biosynthetic cycle. Thus, understanding of the fundamental catalytic mechanisms and protein–protein interactions necessary for polyketide biosynthesis are necessary for successful AT domain specificity engineering. This knowledge, coupled with the advancements outlined in this review, position the AT domain engineer to make huge strides in polyketide diversification in the very near future.

This research was supported by a grant from the National Institutes of Health (no. R01 GM 087934) to C.K.


